The Language Grid: Creating Customized Multilingual Environments

Toru Ishida, Rieko Inaba
Dept. Social Informatics, Kyoto University
NICT Language Grid Project
{ishida, rieko}@i.kyoto-u.ac.jp

Yohei Murakami, Tomohiro Shigenobu
Donghui Lin, Masahiro Tanaka
NICT Language Grid Project
{yohei, shigenobu, lindh, mtnk}@nict.go.jp

Abstract

Since multiple languages are used in various communities in daily life, tools that can effectively support multilingual communication should be provided. We often observe that that the success of a multilingual tool in one situation does not guarantee its success in another. To develop a customized multilingual environment for various situations in various communities, we have implemented the Language Grid; it allows users to freely combine existing language services to develop new services. This paper summarizes the current status of the Language Grid and the lessons learned from this project.

1 Introduction

The Internet allows people to be linked together regardless of location, however language remains the biggest barrier. Its users speak a wide variety of languages. For instance, only 35% of the Internet population speaks English. The remainder is almost equally divided between other European languages and Asian languages. In fact, it is not possible for any one to learn the languages needed to access all possible information from the Internet.

The Language Grid is an infrastructure that is built on the top of the Internet. It allows end users as well as professionals to conquer the language barriers by themselves. Users can combine existing language services provided by researchers and professionals, and create new language services for their own purposes by permitting them to add their own language resources.

Several research groups including NICT (National Institute on Information and Communications Technology), and Kyoto University started working on the Language Grid in April 2006 (Ishida, 2006). This project is based on collaboration between industry, government, universities and NPO/ NGOs. Several practical systems have already been released.

Among existing research studies, EuroWordNet (Vossen, 1998) and Global WordNet Grid (Fellbaum and Vossen, 2007) are pioneer works on connecting dictionaries in different languages based on word semantics. However, the Language Grid is an attempt to build a platform that can combine language services provided by stakeholders with different incentives. Therefore, standardization is quite important (Calzolari 2002).

There also exist several efforts to combine language processing programs: Heart of Gold (Callmeier *et al.*, 2004) and UIMA (Ferrucci and Lally, 2004). They aim at allowing various language processing programs to share data, while the Language Grid is more application oriented and focuses on managing the intellectual property rights associated with language resources based on the service-oriented architecture. Since the motivations are orthogonal, we started bridging Heart of Gold and the Language Grid (Bramantoro *et al.*, 2008), and will apply the results to UIMA.

2 Concept

Online language services already exist including bilingual dictionaries and machine translators. However, difficulties often arise when people try to use those language services in their intercultural activities; as shown in Figure 1, complex contracts, intellectual property rights, and non-standard application interfaces make it difficult for users to create customized language services that support their activities.

To improve the accessibility and usability of existing language services, we need to allow users to easily create new language services by combining existing ones on the Internet. The word “grid” is generally defined as “a system or structure for combining distributed resources; an open standard protocol is generally used to create high quality services.” Our objective, applying the “grid” concept to ensure the collaboration of language services, has not been tried before.

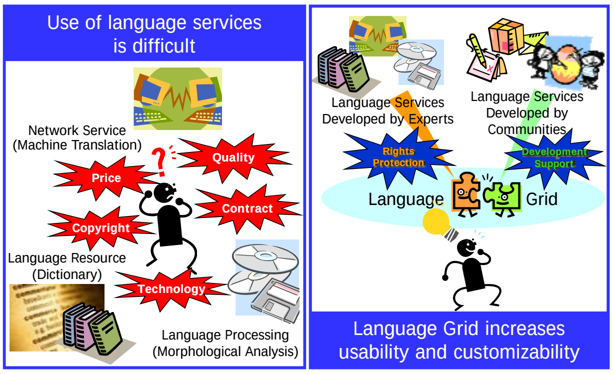

Use of language services is difficult
Network Service (Machine Translation)
Quality
Price
Contract
Copyright
Language Resource (Dictionary)
Technology
Language Processing (Morphological Analysis)
Language Services Developed by Experts
Language Services Developed by Communities
Rights Protection
Development Support
Language
Grid
Language Grid increases usability and customizability

Figure 1: Role of the Language Grid

As shown in Figure 2, since the Language Grid takes the collective intelligence approach, the platform can grow only through the voluntary efforts of users. The more users provide resources, the more fully they can utilize the benefits of the resources. Thus the platform should allow users to create services and share them via the Language Grid. Conceptually, the Language Grid has two main structures: *horizontal* and *vertical*. The horizontal grid concerns the combination of existing bilingual dictionaries or machine translation systems for various languages. The vertical grid concerns specific scenarios of intercultural collaboration activities, which require specialized language services including jargon handling.

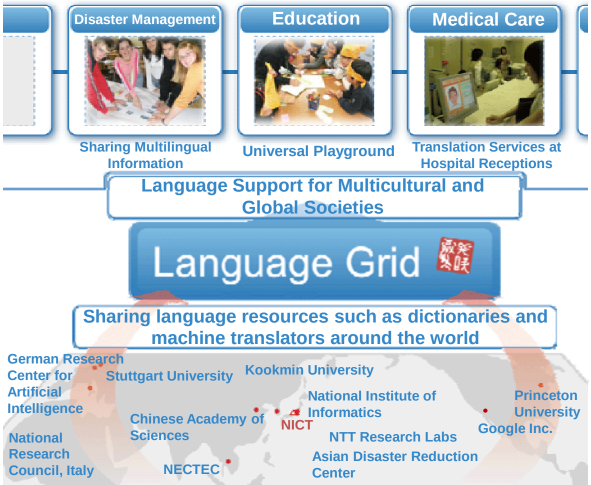

Disaster Management
Education
Medical Care
Sharing Multilingual Information
Universal Playground
Translation Services at Hospital Receptions
Language Support for Multicultural and Global Societies
Language Grid
Sharing language resources such as dictionaries and machine translators around the world
German Research Center for Artificial Intelligence
Stuttgart University
Kookmin University
National Institute of Informatics
Princeton University
Chinese Academy of Sciences
NICT
Google Inc.
NTT Research Labs
National Research Council, Italy
NECTEC
Asian Disaster Reduction Center

Figure 2: The Language Grid Architecture

To realize the Language Grid, Web service technologies including *language service ontology* and *context-aware service compositions* have been developed to enable the collaboration needed among language services. Language service ontology is a technology to define standard language service APIs in a hierarchical way so that end users are provided with simple interfaces while professionals can access more complex interfaces (Hayashi *et al.*, 2008). Context-aware service composition is a technology to combine language services when working on the same document or same conversation: forward and backward translations are coordinated to determine the meanings of words consistently (Rie Tanaka *et al.*, 2009).

3 Service

The Language Grid provides an environment where users can share language resources developed by both professionals and end users in various application fields. Users can register wrap language resources as Web services to register them in the Language Grid. Major stakeholders of the Language Grid fall into three categories.

- Language Grid Operator
  The operator manages the Language Grid and controls language resources and services.
- Language Service Provider
  The service provider registers language services such as machine translations, morphological analyzers, dependency parsers, dictionaries, and parallel texts in the Language Grid.
- Language Service User
  The service user invokes the registered language services for their intercultural activities.

Note that stakeholders are not individuals but groups like research units in universities, and that a single group can act as two different stakeholders: service provider and service user.

As shown in Figure 3, the Language Grid consists of the following service layers.

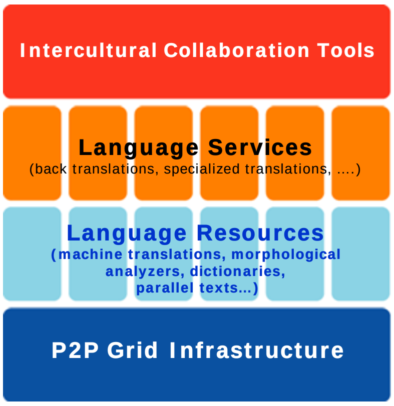

Intercultural Collaboration Tools
Language Services
(back translations, specialized translations, ...)
Language Resources
(machine translations, morphological analyzers, dictionaries, parallel texts...)
P2P Grid Infrastructure

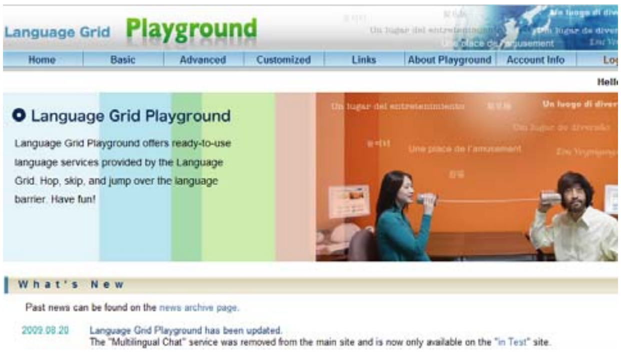
Figure 3: Service Layers

The P2P Grid Infrastructure is aimed at connecting two kinds of servers (*core nodes* and *service nodes*). Core nodes manage all requests to language services, while service nodes actually invoke atomic services. If the requested service is a composite one, core nodes invoke a corresponding Web service workflow that includes one or more atomic services. Registered information of language services is shared among all core nodes. The same services are equally provided, regardless of which core node receives the request. The core nodes also control accesses to services to fulfill service providers' conditions. Using the Language Grid Service Manager (see Figure 6), resource providers can access the usage statistics of the services they provide.

Any user can add new language resources to the Language Grid. A Web service that corresponds to a language resource is called an *atomic service*. To develop atomic services, language resources are wrapped based on language service ontology that standardizes the interfaces of language services.

Atomic language services can be composed by Web service workflows. A service described by a workflow is called a *composite service*. Various composite services have been made available, including back translations and specialized translations. For example, specialized translation can be realized using several atomic services, such as machine translators, morphological analyzers, and domain-specific dictionaries. BPEL4WS is used to describe workflows which are interpreted and executed by a BPEL engine.

Different types of collaboration tools have been developed using the language services explained above. Language Grid Playground[1] provides easy access to the Language Grid to try a variety of registered language services through a Web browser (see Figure 4). Examples of real-world challenges, such as the creation of community dictionaries, or real-world application of the Language Grid technologies, are also introduced through this website.



Language Grid Playground

Language Grid Playground offers ready-to-use language services provided by the Language Grid. Hop, skip, and jump over the language barrier. Have fun!

What's New

Past news can be found on the news archive page.

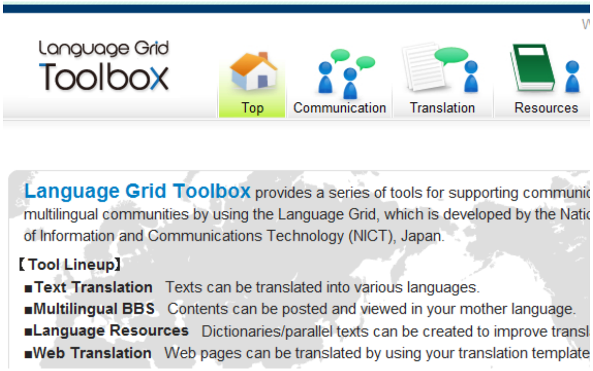
2009.08.20 Language Grid Playground has been updated.
The "Multilingual Chat" service was removed from the main site and is now only available on the "in Test" site.

Figure 4: Language Grid Playground

Services in the Playground are categorized as follows.

- ➢ The BASIC services provide interfaces for easy selection and usage of the atomic language services.
- ➢ The ADVANCED services provide composite language services created by combining existing language services.
- ➢ The CUSTOMIZED services provide specific language services to organizations that carry out intercultural activities.

Language Grid Toolbox[2], on the other hand, is a collection of modules to support multilingual communication in a community (see Figure 5). Users can install this software into their servers to start services, such as multilingual BBS and multilingual dictionary creation. Since Toolbox is based on the open source environment CMS (Content Management System) XOOPS Cube, Toolbox is also provided as open source software. Therefore, the functions of Toolbox can be extended by developing XOOPS modules to meet the requirements of user communities.



Language Grid Toolbox provides a series of tools for supporting communic multilingual communities by using the Language Grid, which is developed by the Natio of Information and Communications Technology (NICT), Japan.

【Tool Lineup】

■Text Translation Texts can be translated into various languages.

■Multilingual BBS Contents can be posted and viewed in your mother language.

■Language Resources Dictionaries/parallel texts can be created to improve transl

■Web Translation Web pages can be translated by using your translation template

Figure 5: Language Grid Toolbox

Furthermore, by using registered language services, existing communication tools can introduce multilingual functions easily. For instance, NOTA [3]and LiquidThreads[4] have been successfully multilingualized.

4 Operation

The Language Grid was developed and released as open source software. Using this source code, universities, research institutes, and NPO/NGOs can develop new tools to support intercultural collaboration or even operate the Language Grid. The Department of Social Informatics at Kyoto University started operation of the Language Grid for nonprofit purposes in December 2007. 116 groups in 17 countries have already joined

[1] http://langrid.org/playground/

[2] http://langrid-tool.nict.go.jp/toolbox/

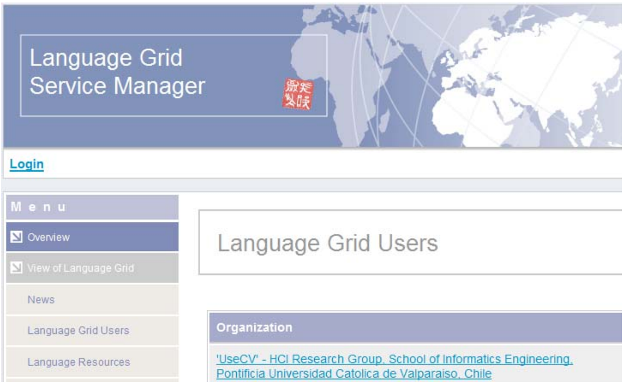
[3] http://notaland.com/

[4] http://www.mediawiki.org/wiki/Extension:LiquidThreads

the Language Grid. Roughly one third of them are providing services, one third of them are using services, and the remaining one third are inactive. We expected that NPO/NGOs and public sectors would become the major users, but universities are using the Language Grid more intensively at this moment.

Research institutes, universities, and companies are providing atomic language services such as dictionaries and machine translators. The number of shared language resources now totals 48. Organizations that provided language resources include Chinese Academy of Sciences, Stuttgart, Princeton, Kookmin, and Kyoto Universities, NICT, National Institute of Informatics, NTT, Google, Toshiba, Oki, Kodensha, Asian Disaster Reduction Center and a number of public sector groups and NPO/NGOs. When providing atomic language services, providers specify copyright notices and license information in the profiles of the resources.

In our operation model (Ishida *et al.*, 2008), language service providers can fully control access to their provided language resources. Language service providers can select users, restrict the total number of accesses per year/ month/day, and set the maximum volume of data transfer per access. Providers set those conditions using the Language Grid Service Manager (see Figure 6).



Language Grid Users

Organization
'UseCV' - HCI Research Group, School of Informatics Engineering, Pontificia Universidad Catolica de Valparaiso, Chile

Figure 6: Language Grid Service Manager

On the other hand, language service users can allow participants in events or activities organized by the users to utilize language services. To avoid the fraudulent usage of language services, however, service users should not allow the participants to discover the ID and password of the Language Grid. For example, in the case of an NPO offering medical interpreter services to foreign patients, the NPO should not enter their Language Grid ID and password in front of the patients, but embed ID and password in their patient support systems.

We have discovered that our current operation model has several drawbacks. The first issue is *centralized operation*. The operation center in Kyoto cannot reach local organizations in other countries. As a result, 74% of participant organizations are in Japan. Since we need global collaboration, even for solving language issues in local communities, this imbalance should be overcome: we need to establish operation centers in different areas and connect them. By introducing the notion of *affiliated operator* and *affiliated user*, we can naturally shift to *federated operation*: affiliated users can use the resources through the affiliated operator. Though there remain several technical issues, we have finalized the new agreement through collaboration with research institutes in Asia and Europe.

Another issue is *non-profit operation*. We are often asked the same question: can the Language Grid be used in the research labs of companies? According to our current operation model, profit organizations can use the Language Grid only for CSR activities, and cannot use it in their research labs. This constraint should be relaxed: university researchers are puzzled by this limitation when working on joint project with companies. Indeed, the original model suppresses many of the potential applications of the Language Grid. Therefore, we will relax this restriction while retaining the control power of service providers. Companies will be encouraged to join the Language Grid, and to provide high quality services, just as Internet service providers do at present.

5 Intercultural Collaboration

At the beginning of the project, users established the Language Grid Association and started activities on intercultural collaboration. Unlike conventional machine translation systems, the Language Grid combines users' language resources and machine translators to produce high-quality translation that can be customized for each field (Ishida, 2010). By utilizing this benefit, NPO/NGOs, schools and other nonprofit sectors have started to play a central role in breaking down the language barriers. Their activities cover a broad range of fields, including disaster management, education, and medical care.

A provider of language services does not have to be a research institute or a university. Organizations that are actually conducting activities for intercultural collaboration can also register their own multilingual services. For example, language resources for schools prepared in various parts of Japan have been wrapped, offered as language services, and shared on the Internet. For example, a language service for medical care is currently used at receptions of Kyoto City Hospital and Kyoto University Hospital.

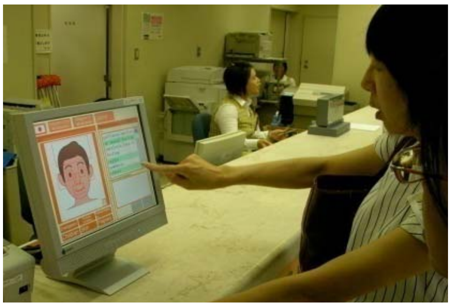
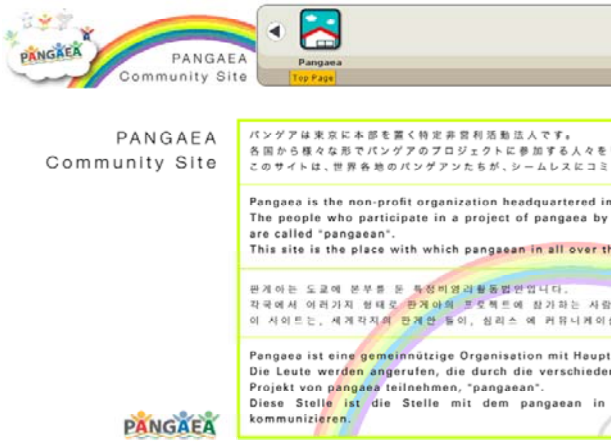


Figure 7: Multilingual Reception System (Provided by Center for Multicultural Society Kyoto and Wakayama University)

The major difference between machine translation on the Language Grid and a conventional translation system on the Internet is that users can themselves improve the quality of translation. First, users can use the registered parallel texts in the translation process. When a user enters a sentence, examples with meanings similar to the entered sentence will appear automatically. If the user can find a sentence that suitably conveys the meaning that he/she intended, he/she can obtain an accurate translation result by using parallel texts. If the user is unable to find the intended expression, machine translation is then executed. In this case, a dictionary registered by the user also helps to improve the quality of translation.



パンゲアは東京に本部を置く特定非営利活動法人です。
各国から様々な形でパンゲアのプロジェクトに参加する人々を
このサイトは、世界各地のパンゲアンたちが、シームレスにコミ

Pangaea is the non-profit organization headquartered in
The people who participate in a project of pangaea by
are called "pangaean".
This site is the place with which pangaean in all over th

판게아는 도쿄에 본부를 둔 특정비영리활동법인입니다.
각국에서 여러가지 형태로 판게아의 프로젝트에 참가하는 사람
이 사이트는, 세계각지의 판게안 들이, 심리스 에 커뮤니케이

Pangaea ist eine gemeinnützige Organisation mit Haupt
Die Leute werden angerufen, die durch die verschiede
Projekt von pangaea teilnehmen, "pangaean".
Diese Stelle ist die Stelle mit dem pangaean in
kommunizieren.

Figure 8: Multilingual Community Site (Provided by NPO Pangaea)

If the quality of translation is not enough, however, another user in the multilingual community might manually correct the translation results. The corrected parallel texts are accumulated so that machine translator can learn from them. This becomes possible when the multilingual community members share their context. NPO Pangaea developed their own community site (see Figure 8) and introduced this process on their multilingual BBS.

6 Research

Various ongoing research activities are using the Language Grid. They are roughly classified into three categories.

The first category is interaction analysis of machine-translation-mediated communication. It appears that machine translations are inconsistent, asymmetric, and intransitive, and thus make it difficult for people to create a conversational common ground (Yamashita and Ishida, 2006) (Yamashita *et al.*, 2009).

The second category is interdisciplinary research using language technologies. Since the Language Grid provides easy-to-use services, researchers who are not linguistic professionals have started to use language technologies (Cho *et al.*, 2008) (Yue Suo *et al.*, 2009).

The third category is developing the Language Grid based on services computing, especially technologies for service composition (Ben Hassine *et al.*, 2006) (Rie Tanaka *et al.*, 2009) and service supervision (Masahiro Tanaka *et al.*, 2009).

7 Conclusion

This paper explained how the Language Grid, an infrastructure that allows end-users to create new language services for their intercultural collaboration activities, increases the accessibility and usability of online language resources. To this end, language resources including data and programs are wrapped as Web services so that users can easily share and combine these Web services for creating their own multilingual environment. Using the Language Grid, various kinds of intercultural activities have begun at hospital receptions, local schools, shopping street communities and so on (Ishida *et al.*, 2007) (Fussell *et al.*, 2009).

This paper also described operation models to coordinate different types of stakeholders in language services. The operation model should be designed to match their incentives. Centralized operation started in December 2007. After two years of operation, we are moving towards federated operation. Further analysis of operations will contribute to increasing the accessibility and usability of language resources.

Acknowledgments

We are grateful to the members of NICT Language Grid Project and Kyoto University Language Grid Operation Center. This research was partially supported by Strategic Information and Communications R&D Promotion Programme from Ministry of Internal Affairs and Communications, and a Grant-in-Aid for Scientific Research (A) (21240014, 2009-2011) from Japan Society for the Promotion of Science (JSPS).

References

Arif Bramantoro, Masahiro Tanaka, Yohei Murakami, Ulrich Schäfer and Toru Ishida. A Hybrid Integrated Architecture for Language Service Composition. *IEEE International Conference on Web Services (ICWS-08)*, pp. 345-352, 2008.

Ulrich Callmeier, Andreas Eisele, Ulrich Schafer and Melanie Siegel. The Deep Thought Core Architecture Framework. *Proceedings of LREC*, pp.1205-1208, 2004.

Nicoletta Calzolari, Antonio Zampolli and Alessandro Lenci. Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. *CICLing*, pp. 264-279, 2002.

Heeryon Cho, Toru Ishida, Toshiyuki Takasaki and Satoshi Oyama. Assisting Pictogram Selection with Semantic Interpretation. *European Semantic Web Conference (ESWC-08)*, Lecture Notes in Computer Science, 5021, Springer-Verlag, pp. 65–79, 2008.

Christiane Fellbaum and Piek Vossen. Connecting the Universal to the Specific: Towards the Global Grid. *Intercultural Collaboration*, Lecture Notes in Computer Science, 4568, Springer-Verlag, pp. 1-16, 2007.

David Ferrucci and Adam Lally. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering, Cambridge University Press*, Vol. 10, pp. 327-348, 2004.

Susan Fussell, Pamela Hinds and Toru Ishida Eds. *The Second International Workshop on Intercultural Collaboration*, ACM Press, 2009

Ahlem Ben Hassine, Matsubara Shigeo and Toru Ishida. Constraint-based Approach for Web Service Composition. *International Semantic Web Conference (ISWC-06)*, Lecture Notes in Computer Science 4273, Springer-Verlag, pp. 130-143, 2006.

Yoshihiko Hayashi, Thierry Declerck, Paul Buitelaar and Monica Monachini. Ontologies for a Global Language Infrastructure. *The First International Conference on Global Interoperability for Language Resources (ICGL-08)*, pp.105-112, 2008.

Toru Ishida. Language Grid: An Infrastructure for Intercultural Collaboration. *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100, 2006.

Toru Ishida. Communicating Culture. *IEEE Intelligent Systems*, Vol. 21, No. 3, pp. 62-63, 2006.

Toru Ishida, S. R. Fussell and P. TJM Vossen Eds. *The First International Workshop on Intercultural Collaboration*, Lecture Notes in Computer Science, 4568, Springer-Verlag, 2007.

Toru Ishida, A. Nadamoto, Y. Murakami, R. Inaba, T. Shigenobu, S. Matsubara, H. Hattori, Y. Kubota, T. Nakaguchi, and E. Tsunokawa. A Non-Profit Operation Model for the Language Grid. *International Conference on Global Interoperability for Language Resources*, pp. 114-121, 2008.

Toru Ishida. Intercultural Collaboration Using Machine Translation. *IEEE Internet Computing*, Vol. 14, No. 1, pp. 30-32, 2010.

Yue Suo, Naoki Miyata, Hiroki Morikawa, Toru Ishida and Yuanchun Shi. Open Smart Classroom: Extensible and Scalable Learning System in Smart Space using Web Service Technology. IEEE Transactions on Knowledge and Data Engineering, Vol.21, No.6, pp. 814-828, 2009.

Masahiro Tanaka, Toru Ishida, Yohei Murakami, and Satoshi Morimoto. Service Supervision: Coordinating Web Services in Open Environment. IEEE International Conference on Web Services (ICWS-09), pp. 238-245, 2009.

Rie Tanaka, Yohei Murakami and Toru Ishida. Context-Based Approach for Pivot Translation Services. *International Joint Conference on Artificial Intelligence (IJCAI-09)*, pp.1555-1561, 2009.

Piek Vossen. Introduction to EuroWordNet. *Computers and the Humanities*, Vol. 32, No. 2-3, pp. 73-89, 1998.

Naomi Yamashita, Reiko Inaba, Hideaki Kuzuoka and Toru Ishida. Difficulties in Establishing Common Ground in Multiparty Groups using Machine Translation. *International Conference on Human Factors in Computing Systems (CHI-09)*, pp. 679-688, 2009.

Naomi Yamashita and Toru Ishida. Effects of Machine Translation on Collaborative Work. *International Conference on Computer Supported Cooperative Work (CSCW-06)*, pp. 515-523, 2006.